A literature review of "DetCon"

Introduction:

Currently large labelled datasets have been supercritical for a supervised Computer Vision task to thrive and succeed. However, in recent times, due to research aimed at reducing the need of labelled data, self-supervised learning models have been able to reduce the need of labels considerably while preserving the efficiency. But these self-supervised models require tremendous computation costs. This crucial bottleneck is superseded by their utility as they allow ways to classify much larger unlabelled datasets. The authors of this paper propose DetCon (Contrastive Detection) an objective which maximizes the similarity in object-level features across augmentation. It achieves this by extracting separate features from each object, hence enriching the information provided by each of the training examples. It also hence provides a more diverse set of negative samples to contrast again. These two points help learn more individual features and hence achieves faster computation and better accuracy than previous attempts. The approximate object based regions are identified through unsupervised segmentation algorithms. Perceptual grouping is one such prior used which performs self supervised image segmentation through object features such as colour, texture, orientation and other mid level features. DetCon was used in pre-training to learn transferable features quickly. It was weighed in popular supervised algorithms and other self supervised algorithms like SIMCLR and BYOL. Most of the other recent methods been able to impressively increase the unsupervised accuracy but fail to increase the pre-training efficiency. DetCon is hence able to alleviate the computation burden by upto 10x the computation for supervised transfer learning on ImageNet. Hence as a result, longer training periods also lead to state-of-the-art transfer to COCO(common objects in context dataset) detection and instance segmentation. Authors also claim to have on par accuracy on state-of-the-art self-supervised system SEER which has 1000 times more input images.

Early work in the field of self supervised pre-training focussed on improving pre-training architecture and data. Early self-supervised pre-training involved image restoration tasks with a few methods to study higher level pretexts tasks as pre-training. Contrastive objectives which minimize similarity in opposite images and increase similarity in same images are recently becoming more popular. While most work is more inclined towards learning representations of the whole image, few works have been trying to learn local descriptors that are more relevant for downstream tasks such as detection and segmentation. There were a few works which are similar to learning from local descriptors with discriminating changes from DetCon being the specialized architecture backbones for segmentation and the use of different loss functions. Although these works have been able to achieve impressive accuracy, they seem to not report any gains in terms of pre-training efficiency for transfer learning tasks such as on COCO detection.

The Approach:

The process starts with random augmentation(like random cropping, flipping, etc) on the image hence providing us with 2 new images. In addition to this, the authors have also created a segment mask which divides the segment into different components. These masks were computed using unsupervised segmentation algorithms. These masks are also augmented corresponding to the underlying image to get 2 sets of masks as well which are aligned to the 2 augmented images. Then, the images are encoded with ResNet-50 encoder into a spatial map of hidden vector representations of the images. Every mask m associated with the image, it is also pooled down into a mask-pooled hidden vector using average pooling. They are then fed into a two layer MLP for inducing non-linearity. The encoder network is the network which is used further for transfer learning. The DetCon used with SIMCLR paradigm has 2 views trained through the same encoder

and projection network whereas the DetCon with BYOL paradigm has one view with the projection and encoder network with weights as *theta* and the other view with weights as the exponential moving average of *theta.* Then the views are then used to calculate the loss through the novel contrastive detection loss function.

The algorithm works in such a way that it helps optimize the feature extraction by instilling the masked pooling and the contrastive detection loss which increases the dissimilarity between different masks in the same image. This objective function is then specially curated by sampling 16 masks(possibly redundant) at every iteration. Then the similarity between each image and mask pair is calculated. Then the similarity between paired locations is maximised. This also handles the case where the mask is present in one view and not in the other. That is we first check if the 2 masks correspond to the same underlying region and then minimize the dissimilarity between the views by the corresponding factor.

Evaluation and Inference:

The computational cost associated with the algorithm is mainly because of the ResNet50 encoder. The extra computation also arises from the projection heads and the 16 hidden vectors forwarded through these projection heads instead of 1. The Segmentation does not take a lot of computation as once the masks have been created, they can simply be reused throughout the training process. Different Image segmentation techniques such as spatial heuristics, image computable and human segmentation masks were used and here lies scope to study how different segmentation techniques affect the results.

The model was pre-trained on ImageNet and fine-tuned and evaluated on COCO for object detection and image segmentation, semantic segmentation on PASCAL and Cityscapes and depth estimation on NYU v2 datasets. It required upto 10x less pre-training and provided with an efficiency gain of about 3x to 10x on different tasks when compared to the original algorithm of BYOL or SIMCLR. DetCon outperforms all the other prior arts by a considerable margin. Also evaluating what happens when larger models are used for example ResNet 200 or ResNet 152 instead of ResNet50, DetCon continues to outperform its competitors.

DetCon has shown that self-supervised learning approaches do not need many negative samples, instead something similar to the attention mechanism from a transformer to differentiate between a single object and its surroundings. This poses a significant question about the need for distinct features and objects in the image. The authors have not provided results when the objects are not distinct or if multiple objects in an image contribute to the final classification of the image.

Also, as the authors have told in the publication, the better the mask covers the ground truth object, the better DetCon performs. However, looking at the self supervised semantic segmentation algorithms used, there is a major chance of improvement. If the masking is done using a semi-supervised approach(example semi-supervised algorithm similar to the pseudo-labelling performed with Mask-RCNN), the masks can more effectively cover the ground truth objects and the training efficiency can be increased with the small trade-off of human annotating the data.

Apart from these few caveats, the DetCon outperforms its competitors by a large margin and has introduced a novel idea of using semantic segmentation for self-supervised learning. There is a future scope for trying different ideas along with it. There also lies a prospect of using transformers for the semantic segmentation using patches instead of masks. The DetCon has surely set up a plethora for new research to build upon.